Country Corruption Analysis with Self Organizing Maps and Support Vector Machines

Johan Huysmans[1], David Martens[1], Bart Baesens[2,1], Jan Vanthienen[1], and Tony Van Gestel[3,4]

[1] Department of Decision Sciences and Information Management, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium
[2] School of Management, University of Southampton, Southampton, SO17 1BJ, United Kingdom
[3] Credit Risk Modelling, Dexia Group, Square Meeus 1, B-1000 Brussels, Belgium
[4] Department of Electrical Engineering, ESAT-SCD-SISTA, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium

Abstract. During recent years, the empirical research on corruption has grown considerably. Possible links between government corruption and terrorism have attracted an increasing interest in this research field. Most of the existing literature discusses the topic from a socio-economical perspective and only few studies tackle this research field from a data mining point of view. In this paper, we apply data mining techniques onto a cross-country database linking macro-economical variables to perceived levels of corruption. In the first part, self organizing maps are applied to study the interconnections between these variables. Afterwards, support vector machines are trained on part of the data and used to forecast corruption for other countries. Large deviations for specific countries between these models' predictions and the actual values can prove useful for further research. Finally, projection of the forecasts onto a self organizing map allows a detailed comparison between the different models' behavior.

1 Introduction

The amount of empirical research concerning corruption is impressive. An overview can be found in [1, 2]. Most of existing literature tries to find causal relations between some explanatory variable and the perceived level of corruption. For example, in [3] the influence of democracy on the perceived level of corruption is tested while other studies focus on the influence of religion [4], colonial heritage [4], abundance of natural resources [5] or the presence of women in parliament [6]. Other studies focus on the consequences of corruption: does corruption lead to a decrease of GDP, foreign investments or aid [7]? The main problem in all these empirical studies is to make the transition from 'highly correlated' to 'causes': many variables are highly correlated with the perceived level of corruption, but it is difficult to derive causal relations from it. For example, in [2] is reported that

H. Chen et al. (Eds.): WISI 2006, LNCS 3917, pp. 103–114, 2006.

GDP per head and corruption are reported to be highly positively correlated in most studies but that there is general agreement that there is no causality involved.

The general approach in the majority of these studies (e.g., [3]) is to regress a variable representing corruption on a number of independent variables for which the influence is tested with the possible inclusion of some control variables. In this paper, we apply a different technique to study corruption. We use self organizing maps (SOMs), also known as Kohonen maps, to gain deeper insight in the causes of corruption. This technique is derived from the data mining community and allows a clear and intuitive visualization of high-dimensional data. In the second part of the paper, we apply support vector machines (SVMs) to forecast changes in the perceived levels of corruption. Support vector machines have proven to be excellent classifiers in other application domains (e.g., credit scoring [8]) and are able to capture nonlinear relationships between the dependent and independent variables.

In the next section, a short introduction to the concept of SOMs is given. Afterwards, we describe the data that was used for this study. The next section discusses the application of SOMs on this data whereby special attention is paid to the visualization possibilities that these models offer. In the final section, we use least-squares support vector machines [9] to forecast changes in the perceived level of corruption. Input selection will be used to select the most significant variables. The main contribution of this paper is the projection of SVM predictions onto a SOM to gain more insight in the SVM model and the use of multi-year data sets to study evolutions in the perceived level of corruption.

2 Self Organizing Maps

SOMs were introduced in 1982 by Teuvo Kohonen [10] and have been used in a wide array of applications like the visualization of high-dimensional data [11], clustering of text documents [12], identification of fraudulent insurance claims [13] and many others. An extensive overview of successful applications can be found in [14] and [15]. A SOM is a feedforward neural network consisting of two layers. The neurons from the output layer are usually ordered in a low-dimensional grid. Each unit in the input layer is connected to all neurons in the output layer with weights attached to each of these connections. This is similar to a weight vector, with the dimensionality of the input space, being associated with each output neuron. When a training vector $\boldsymbol{x}$ is presented, the weight vector of each neuron c is compared with $\boldsymbol{x}$. One commonly opts for the euclidian distance between both vectors as the distance measure. The neuron that lies closest to $\boldsymbol{x}$ is called the 'winner' or the Best Matching Unit (BMU). The weight vector of the BMU and its neighbors in the grid are adapted with the following learning rule:

$$\boldsymbol{w}_c = \boldsymbol{w}_c + \eta(t)\Lambda_{winner,c}(t)(\boldsymbol{x} - \boldsymbol{w}_c) \tag{1}$$

In this expression $\eta(t)$ represents the learning rate that decreases during training. $\Lambda_{winner,c}(t)$ is the so-called neighborhood function that decreases when the distance in the grid between neuron c and the winner unit becomes larger. Often a gaussian function centered around the winner unit is used as the neighborhood function with a decreasing radius during training. The decreasing learning rate and radius of the neighborhood function result in a stable map that does not change substantially after a certain amount of training.

From the learning rule, it can be seen that the neurons will move towards the input vector and that the magnitude of the update is determined by the neighborhood function. Because units that are close to each other in the grid will receive similar updates, the weights of these neurons will resemble each other and the neurons will be activated by similar input patterns. The winner units for similar input vectors are mostly close to each other and self organizing maps are therefore often called topology-preserving maps.

3 Description and Preprocessing of the Data

For this study, data from three different sources was combined. Demographic information, for example literacy and infant mortality rate, was retrieved from the CIA Factbook [16] together with macro-economical variables, like GDP per capita and sectorial GDP information.

Information concerning the corruption level in specific countries was derived from Transparency International [17] under the form of the Corruption Perceptions Index (CPI). This index ranks countries according to the degree to which corruption is perceived to exist among public officials and politicians. The CPI is created by interviewing business people and country analysts and gives a score between 0 (highly corrupt) and 10 (highly clean) to each country. In this study, data concerning the years 1996, 2000 and 2004 was used. In the index of 1996, 54 countries received a corruption score. We select only these countries and omit from the more elaborated 2000 and 2004 indices all other countries, resulting in a total of 162 observations: three observations from different years for each of the 54 countries[1]. We use ISO 3166 Codes, like BEL for Belgium or FRA for France, to refer to the individual countries whereby capitalization is used to indicate the year of the observation. Uppercase codes (e.g., BEL) indicate that the observation is from 2004, lowercase codes (e.g., bel) are used to refer to observations from the year 2000 and codes in proper case (only the first letter capitalized e.g., Bel) refer to 1996 observations.

Information about the democracy level in each country was obtained from Freedom House [18]. Each country is assigned a rating for political rights and a rating for civil liberties based on a scale of 1 to 7, with 1 representing the highest degree of freedom present and seven the lowest level of freedom. Similarly to the 'level of corruption', the 'level of democracy' in a country is a rather subjective

[1] Pakistan and Bangladesh received a CPI rating in 1996 and 2004, but not in 2000. These two countries are not removed from the data set.

Table 1. Variables included in the dataset

Corruption Perceptions Index (CPI)
Civil Liberties (CL)
Political Rights (PR)
Arable Land (%)
Age structure: 0-14 years
Age structure: 15-64 years
Age structure: 65 years or over
Population growth rate
Birth rate
Death rate
Net migration rate
Total Infant mortality rate (IMR)
Total Life expectancy at birth
Total fertility rate (children born/women)
GDP per capita
GDP agriculture
GDP industry
GDP services
Number of international organisations the country is member of
Literacy(Total Population %)

concept and therefore difficult to express in only two indices. We refer to [3] for some critiques on the inclusion of these indices.

An overview of all variables that were used in this study is given in Table 1.

4 SOM Analysis

4.1 Exploring the Data

We started by training a self-organizing map of 15 by 15 neurons: with this size it can be expected that each neuron will be the BMU for at most a few observations and this allows a clear visualization of the map. All available variables, including the CPI-scores, were used to create the map of Figure 1. We can see that European countries are likely to be projected on the upper right corner of this map, while the other corners are dominated by respectively African, South-American and Asian countries. A second point that draws the attention is the fact that for most countries the observations from the three years lie close to each other. The observations are either projected on the same neuron (e.g., Uganda (UGA), Denmark (DNK), Hong Kong (HKG)) or on adjacent neurons (e.g., Mexico (MEX), Australia (AUS)). The map also shows that for most of the countries for which the position changed over time, the capitalized 2004 observation lies closer to the upper right corner than the 1996 and 2000 observations. This is the case for Mexico (MEX), Brazil (BRA), Thailand (THA), Argentina (ARG), Chile (CHL) and Ecuador (ECU). It seems that these countries are in transition towards a "more European" model.

While the map of Figure 1 provides general indications about the degree of similarity between countries, it does not allow us to obtain detailed information about corruption. To overcome this limitation, component planes were used to gain deeper insight in the data. Component planes can be created for each input variable and

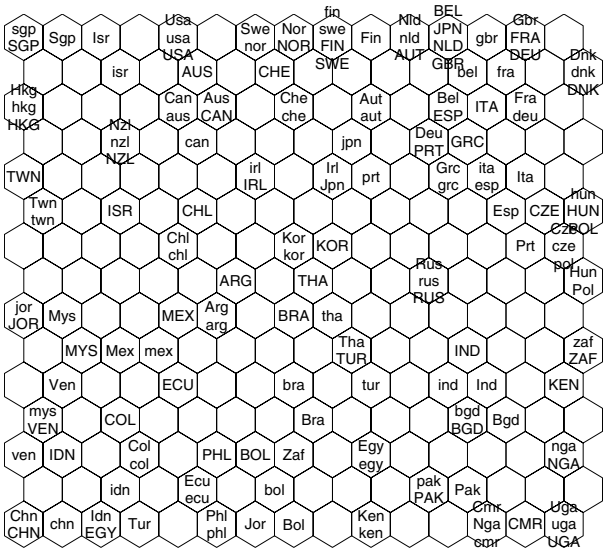


Fig. 1. Map of 15 by 15 neurons

show the weights that connect each neuron with the particular input variable. The component plane for the Corruption Perceptions Index is shown in Figure 2. In this Figure, light and dark shades indicate respectively 'non corrupt' and 'highly corrupt' countries. We can observe that the lower right corner contains the countries perceived to be most corrupt (e.g., Pakistan (PAK), Nigeria (NIG), Cameroon (CMR) and Bangladesh (BGD)). At the opposite side, it can easily be noted that the North-European countries are perceived to be among the least corrupt: they are all situated in the white-colored region at the top of the map. Remember that most European countries were also projected on the upper-half of the map indicating a modest amount of corruption and that several countries seemed to be in transition towards a more European- less corrupt- model.

Component planes for other variables are shown in Figure 3. The first component plane provides information about the literacy of the population. The dark spot indicates the countries where most of the population is illiterate. The second component plane shows Freedom House's index of Political Rights. The light colored spots indicate the regions on the map with the countries that score low on 'political freedom'. The resemblance between these two component planes and the component plane of the corruption index is remarkable. There is a significant correlation between 'corruption', 'literacy' and 'political freedom'. The third component plane (Figure 3(c)) shows the number of international organizations that each country is member of. This component plane can be used to test the hypothesis that corrupt countries are less likely to be member of international organizations because they are either not welcome or not willing to participate. We can see that this hypothesis can not be confirmed based on the

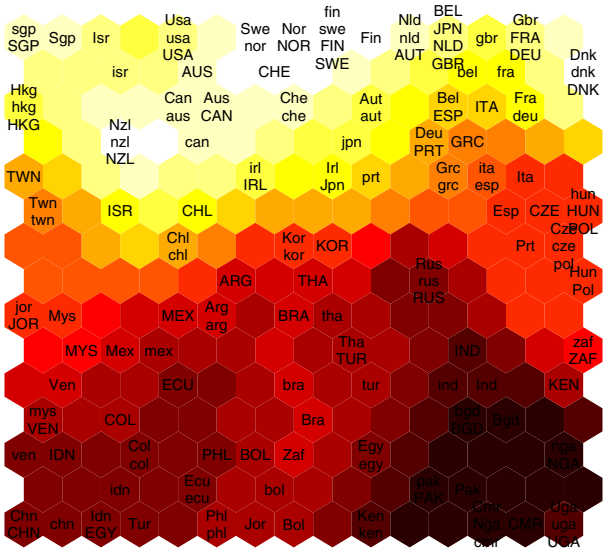
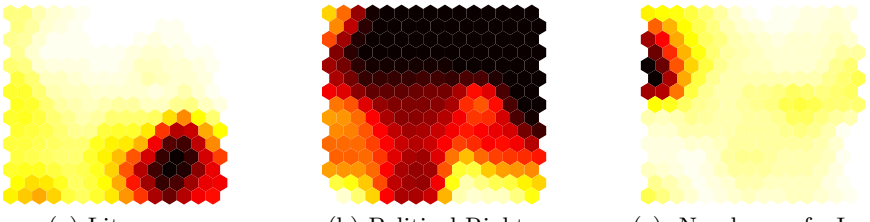


Fig. 2. Corruption Perceptions Index



(a) Literacy

(b) Political Rights

(c) Number of International Organizations the Country is Member Of

Fig. 3. Component Planes

component plane. Countries in the corrupt region of the map do not differ from most European countries. Only some countries or regions close to the upper left corner (Hong Kong (HKG) and Taiwan (TWN)) seem to participate in fewer international organizations.

The same kind of analysis can be performed for each of the other input variables. Several of these component planes, like 'GDP per capita' or 'Total Life Expectancy', show a high degree of correlation with the CPI component plane. Others, like 'Birth Rate' or '% GDP in agricultural sector' are inversely correlated with the CPI component plane, indicating that in those countries corruption goes hand in hand with a young population and little industrialization.

4.2 Clustering of the SOM

In the preceding section, we have shown that SOMs are a suitable tool for the exploration of data sets. If the SOM grid itself consists of numerous neurons, analysis can be facilitated by clustering similar neurons into groups [19].

We performed the traditional k-means clustering algorithm on the trained map of 15 by 15 neurons. The result of this procedure, assuming 5 clusters are present in the data, is given in Figure 4. Several unsupervised labelling techniques have been proposed to identify those variables that characterise a certain cluster (e.g., [20–22]). We briefly discuss the method presented in [20].

After performing a clustering of the trained map, the unlabelled training examples are projected again onto the map and it is remembered which observations are assigned to each cluster. Consequently, for each cluster the so-called salient dimensions are sought. These are defined as the variables that have significantly different values for the observations belonging and not-belonging to that cluster. Finally, a human domain specialist can manually interpret the salient dimensions of each cluster and assign a descriptive label.[2]

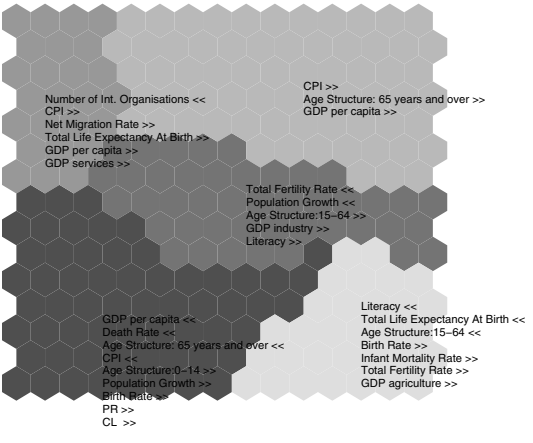


Fig. 4. Labelled Clusters

The above algorithm was executed on the clusters found by the k-means algorithm and the results are also shown in Figure 4. In each cluster we show the variables that were found to be salient. A >> (<<) sign behind the variable

[2] We have made one important change to the algorithm described in [20]. The original algorithm calculates difference factors based on the following formula $\frac{\mu_{in}(k,v)-\mu_{out}(k,v)}{\mu_{out}(k,v)}$ and uses these difference factors to find the salient dimensions (=dimensions with large or small difference factors). This has the disadvantage that for $\mu_{out}(k,v)$ close to zero the difference factor becomes very large even if the difference between $\mu_{out}(k,v)$ and $\mu_{in}(k,v)$ is small. To avoid this problem we add 1 in both numerator and denominator when calculating the difference factors.

name indicates that countries within the cluster have on average significantly larger (smaller) values for this variable than countries outside the cluster. We can see that the analysis of the component planes can be largely automated by this procedure. For example, the procedure clearly indicates that countries from the cluster in the upper left corner seem to participate in few international organizations and score high on the CPI (low corruption). The cluster in the lower right corner on the other hand, is characterized by a low level of literacy, high fertility rate and a large percentage of GDP obtained from agriculture. A 'high level of corruption' is not preserved as a salient characteristic of this cluster because only the most salient features were included on this map and features like 'literacy' and 'agricultural GDP' were found to be more significant.

5 SVM Analysis

In this section several linear and nonlinear corruption forecasting models are evaluated. The models that are derived can be used for different goals. For countries without a CPI-score, the models can be adopted to obtain an estimate of the corruption in that country. For countries that have received a CPI-score, we can apply the models to investigate which countries have a CPI-score that is significantly different from the model predictions and study the reasons for these deviations.

In this part of the paper, the information that is available in the data of 1996 is used to make forecasts of the corruption level in 2000 and 2004. Thus, we use the 54 observations of 1996 for training a LS-SVM model and use this model to predict the corruption levels in 2000 and 2004. Additionally, the data sets of 2000 and 2004 are expanded by including those countries for which a 2000 or 2004 CPI value was available but a 1996 CPI value was not available. The result is a test data set consisting of 231 countries with 125 of them being observations from countries that are not present in the 1996 training data set. We will refer to these 125 observations as the 'new' (unseen) test data, while the other 106 observations are referred to as 'old' (unseen, but an observation for the same country is present in the training data).

First, a linear OLS regression model was constructed whereby input selection was performed on the variables of Table 1. Feature selection was performed with a backwards procedure, but instead of using the traditional t-statistics to decide which variable to remove, we use the leave-one-out crossvalidation error on the training data set. The complete feature selection procedure occurs as follows:

1. Build a linear model containing all available inputs and measure the leave-one-out error.
2. Remove each of the variables one at a time and measure the leave-one-out error of the resulting model.
3. Select the model with the smallest leave-one-out error and go back to the previous step until a model is obtained with only one variable left.

Afterwards, one selects from all models created in step 3 the model with the smallest overall leave-one-out error. This model is then used to create predictions for the 2000 and 2004 test observations.

If the above procedure is performed on the corruption data sets, the model with 5 variables is selected as the best performing model. An overview of the model (with corresponding t-statistics between brackets) is given in Equation 2:

$$\begin{aligned} CPI = 3.005 &+ \underset{(2.4399)}{1.0666 * 10^{-4}\ GDP \text{ per capita}} - \underset{(-3.3354)}{0.88722\ CL} - \underset{(-1.5189)}{0.015074\ IMR} \\ &+ \underset{(1.9103)}{0.40462\ PR} + \underset{(2.0301)}{0.052143\ GDP \text{ services}} \end{aligned} \tag{2}$$

Observe that the signs of the equation parameters correspond mostly to what one could expect based on common sense. Increases in 'GDP per capita' and 'GDP services' increase the CPI score (corresponding to a decrease in corruption) and vice versa for the 'Infant Mortality Rate'. The two variables approximating the democracy level in a country, Civil Liberties (CL) and Political Rights (PR), are also among the features selected by the backwards procedure. The negative sign for the CL-index indicates that an increase of the CL-index (less freedom) decreases the CPI-index (higher corruption). The positive sign of the PR-index seems counter-intuitive: an increase of the PR-index (less political rights) results in an increase of the CPI index (less corruption). This result might however be explained by research from Montinola and Jackman [23]. They found that the level of corruption is typically lower in dictatorships than in countries that have partially democratized, but once past a threshold, democratic practices inhibit corruption. If this non-linearity is indeed present, we should observe significant performance improvements when using a nonlinear LS-SVM model. This LS-SVM model is trained on the same five variables and suitable parameters for regularization and the RBF-kernel are selected by a gridsearch procedure as described in [9].

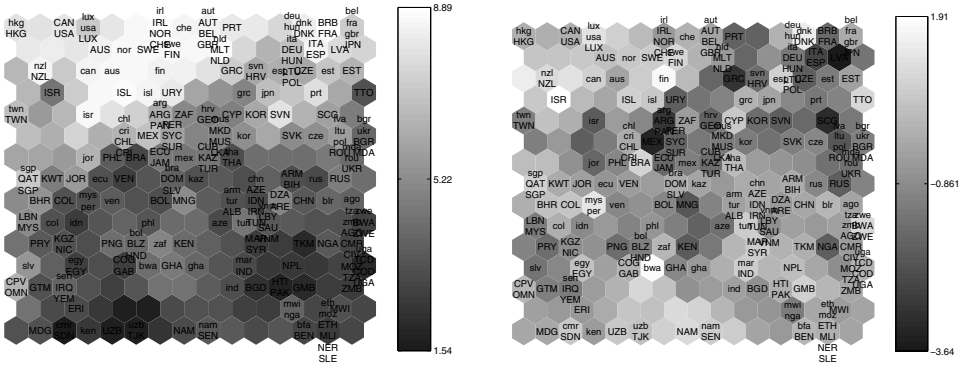
The constructed models are tested on the remaining data of 2000 and 2004. The results are shown in Table 2. It can be observed from the large R^2s that both models are able to explain most of the variance in the corruption index. The small values for the Mean Absolute Error (MAE) confirm this: on average the predictions differ only 0.86 and 0.78 units from the actual observed values for both models on the training data. The results on the test data are also shown in Table 2, where overall performance is indicated together with a breakdown by category. While the MAE is similar for 'old' and 'new' test data, there are huge differences in R^2. The main reason for this deviation is due to the fact that the observations from the 'new' test data have a smaller mean and variance of the CPI than observations from the 'old' test data, indicating that countries that were added to the CPI in recent years are on average more corrupt.

Table 2. Overview Model Performance

	Linear Model		LS-SVM Model	
	R^2	MAE	R^2	MAE
Training Data	0.81	0.86	0.84	0.78
Test Data	0.67	1.07	0.71	0.96
New	0.34	1.05	0.42	0.97
Old	0.73	1.09	0.75	0.98

6 Integration of SOM and SVM

From Table 2, one can observe that the LS-SVM model provides a better forecasting accuracy than the corresponding linear model. However, the LS-SVM has a serious disadvantage: it is very difficult to understand the motivation behind this model's decisions due to its complexity. To relieve this opacity restriction and to gain more insight in the model's behavior we will project its forecasts and the forecasting errors onto a self organizing map (Figure 5).



(a) LS-SVM predictions (b) LS-SVM errors

Fig. 5. Projection of Support Vector Machine onto Self Organizing Map

To create Figure 5(a), a SOM is constructed from the 1996 training data, without use of the CPI, and on this trained map the observations from the test data are projected. Afterwards, each neuron of this map is assigned a color based on the average LS-SVM forecasts of the test observations projected onto that neuron. The neurons that were never the BMU are assigned the average color of the surrounding neurons. The same method was used to create Figure 5(b), but with the colors based on the forecasting errors instead of the actual forecasts. From Figure 5(a), it can be observed that observations projected on the upper half of the map are predicted to be among the least corrupt countries. From Figure 5(b), we learn that the model errors are evenly divided over the

map. There are no regions where the LS-SVM model systematically over- or underestimates the perceived level of corruption. Both figures were also created for the linear model (not shown) and this allowed us to find out why this model was performing worse. The linear model's forecasts were very similar to the ones of the LS-SVM model, except for the lower right corner. In this region the linear model was systematically overestimating corruption, i.e. the actual corruption was less than predicted. Visual inspection also learned that both linear and LS-SVM model made similar forecasting errors for particular observations. For ARG (Argentina), GRC (Greece) and MEX (Mexico) the actual level of corruption is significantly higher than predicted by both models while the opposite is valid for bwa (Botswana) and fin (Finland). Further research is necessary to reveal the reasons for these deviations.

7 Conclusion

In this paper, a data mining approach for the analysis of corruption was presented. In the first part, the powerful visualization possibilities of self organizing maps were used to study the interconnections between various macro-economical variables and the perceived level of corruption The use of multi-year data sets allowed us to visualize the evolution of corruption over time for specific countries. In the second part, it was shown that forecasting models can be constructed that allow analysts to predict the level of corruption for countries where this information is missing. Finally, it was shown how self organizing maps can be used to study the behavior of supervised models. This allows us to open the 'black box' of some models and a more detailed comparison between the predictions made by different models.

References

1. Gerring, J., Thacker, S.: political institutions and corruption: The role of unitarism and parliamentarism. The British Journal of Political Science **34** (2004) 295–330
2. Lambsdorff, J.: Corruption in empirical research: a review. Transparency International Working paper (1999)
3. Bohara, A., Mitchell, N., Mittendorff, C.: Compound democracy and the control of corruption: A cross-country investigation. The Policy Studies Journal **32**(4) (2004) 481–499
4. Treisman, D.: The causes of corruption: a cross-national study. Journal of Public Economics **76**(3) (2000) 339–457
5. Leite, C., Weidmann, J.: Does mother nature corrupt? natural resources, corruption and economical growth. International Monetary Fund Working Paper 99/85 (1999)
6. Swamy, A., Knack, S., Lee, Y., Azfar, O.: Gender and corruption. Journal of Development Economics **64** (2001) 25–55
7. Alesina, A., Weder, B.: Do corrupt governments receive less foreign aid? National Bureau of Economic Research Working Paper 7108 (1999)

8. Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J.: Benchmarking state of the art classification algorithms for credit scoring. Journal of the Operational Research Society **54**(6) (2003) 627–635
9. Suykens, J., Gestel, T.V., Brabanter, J.D., Moor, B.D., Vandewalle, J.: Least Squares Support Vector Machines. World Scientific, Singapore (2002)
10. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics **43** (1982) 59–69
11. Vesanto, J.: Som-based data visualization methods. Intelligent Data Analysis **3** (1999) 111–26
12. Honkela, T., Kaski, S., Lagus, K., Kohonen, T.: WEBSOM—self-organizing maps of document collections. In: Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4-6. Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland (1997) 310–315
13. Brockett, P., Xia, X., Derrig, R.: Using kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. International Journal of Risk and Insurance **65** (1998) 245–274
14. Kohonen, T.: Self-Organising Maps. Springer-Verlag (1995)
15. Deboeck, G., Kohonen, T.: Visual Explorations in Finance with selforganizing maps. Springer-Verlag (1998)
16. CIA: (http://www.cia.gov/cia/publications/factbook/)
17. Transparency International: (http://www.transparency.org/)
18. Freedom House: Freedom in the world country ratings (2005)
19. Vesanto, J., Alhoniemi, E.: Clustering of the self-organizing map. IEEE Transactions on Neural Networks **11**(3) (2000) 586–600
20. Azcarraga, A., Hsieh, M., Pan, S., Setiono, R.: Extracting salient dimensions for automatic som labeling. Transactions on Systems, Management and Cybernetics, Part C **35**(4) (2005) 595–600
21. Lagus, K., Kaski, S.: Keyword selection method for characterizing text document maps. In: Proceedings of ICANN99, Ninth International Conference on Artificial Neural Networks. IEE (1999) 371–376
22. Rauber, A., Merkl, D.: Automatic labeling of self-organizing maps: Making a treasure-map reveal its secrets. In: Proceedings of the third Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'99). LNCS / LNAI 1574, Springer Verlag (1999) 228–237
23. Montinola, G., Jackman, R.: Sources of corruption: a cross-country study. British Journal of Political Science **32** (2002) 147–170